nature genetics

Independent specialization of the human and mouse X chromosomes for the male germ line

Jacob L Mueller[1], Helen Skaletsky[1,2], Laura G Brown[1,2], Sara Zaghlul[1], Susan Rock[3], Tina Graves[3], Katherine Auger[4], Wesley C Warren[3], Richard K Wilson[3] & David C Page[1,2,5]

We compared the human and mouse X chromosomes to systematically test Ohno's law, which states that the gene content of X chromosomes is conserved across placental mammals[1]. First, we improved the accuracy of the human X-chromosome reference sequence through single-haplotype sequencing of ampliconic regions. The new sequence closed gaps in the reference sequence, corrected previously misassembled regions and identified new palindromic amplicons. Our subsequent analysis led us to conclude that the evolution of human and mouse X chromosomes was bimodal. In accord with Ohno's law, 94–95% of X-linked single-copy genes are shared by humans and mice; most are expressed in both sexes. Notably, most X-ampliconic genes are exceptions to Ohno's law: only 31% of human and 22% of mouse X-ampliconic genes had orthologs in the other species. X-ampliconic genes are expressed predominantly in testicular germ cells, and many were independently acquired since divergence from the common ancestor of humans and mice, specializing portions of their X chromosomes for sperm production.

In 1967, Susumu Ohno predicted that catalogs of X-linked genes would differ little, if at all, among placental mammals[1]. Over the past 15 years, numerous comparative mapping studies across highly diverged mammals have supported what has become known as Ohno's law[2–11], although some individual gene exceptions have been noted[12,13]. We decided to perform a systematic and rigorous test of Ohno's law by comparing the human and mouse X chromosomes, including their gene contents. We chose these two X chromosomes because their reference sequences were determined via a high-quality, clone-based approach[14], were verified with high-resolution genetic maps[15,16], harbor substantially fewer gaps than all other sequenced X chromosomes (**Table 1**) and have been well annotated[5,17].

A major difference between these two assemblies is that the mouse X-chromosome reference assembly is derived from a single haplotype[17], whereas the human X-chromosome reference represents a mosaic of X-chromosome sequences from at least 16 different individuals[5]. This mosaicism could have led to misassemblies in the human X-chromosome reference sequence, which, if left uncorrected, would confound our thorough testing of Ohno's law. Such misassemblies might explain why the human X-chromosome reference sequence does not contain the seven large ampliconic regions (segmental duplications of >10 kb in length that share >99% nucleotide identity) found in the mouse X-chromosome reference sequence (**Fig. 1**). Ampliconic regions are particularly prone to sequence misassembly[14] because the nucleotide identity of two amplicons (99.02–99.98%) is comparable to if not greater than the nucleotide identity of alleles (which can be as low as 99.40%; ref. 18). Ampliconic regions assembled from multiple haplotypes may have expansions, contractions or inversions that do not accurately reflect the structure of any extant haplotype. To thoroughly test Ohno's law, we constructed a more accurate assembly of the human X chromosome's ampliconic regions to compare the gene contents of the human and mouse X chromosomes.

We first identified all ampliconic regions of the human X chromosome, including those absent from the current reference sequence. We found 24 ampliconic regions present in the reference sequence by searching for duplicated segments of >10 kb in length and exhibiting >99% nucleotide identity. To identify amplicons absent from the current reference sequence, we targeted regions surrounding gaps, which are generally enriched for amplicons[19], and regions where the reference sequence was discordant with a set of independent physical maps[20]. Together, these approaches yielded a total of 33 regions that merited scrutiny (**Supplementary Table 1**). Only 4 of the 33 regions were spanned by single-haplotype sequence, highlighting the mosaic nature of the human X-chromosome assembly. We chose to resequence the other 29 regions using an approach previously developed by our laboratories to sequence Y-chromosomal amplicons: single-haplotype iterative mapping and sequencing (SHIMS)[21–24]. This clone-based sequencing strategy uses single-nucleotide differences between overlapping clones, all derived from a single haplotype, to accurately order and orient each clone across ampliconic sequences.

Using SHIMS, we generated 11.5 Mb of non-overlapping sequence from 110 newly sequenced BACs, 28 reassembled BACs and 13 fosmids that collectively spanned all 29 regions (**Supplementary Table 1**). Of the

[1]Whitehead Institute, Cambridge, Massachusetts, USA. [2]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]The Genome Institute, Washington University School of Medicine, St. Louis, Missouri, USA. [4]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [5]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Correspondence should be addressed to D.C.P. (dcpage@wi.mit.edu).

Received 11 February; accepted 20 June; published online 21 July 2013; doi:10.1038/ng.2705

Table 1 X-chromosome sequence assemblies in placental mammals

Organism	Sequencing strategy	Number of gaps in X-chromosome assembly	X-ampliconic sequence (Mb)	Percent of X chromosome containing amplicons
Human	Clone based	5	3.15	2.0
Mouse	Clone based	25	19.42	11.6
Chimpanzee	Whole-genome shotgun	10,286	0.00[a]	0.00[a]
Rhesus	Whole-genome shotgun	1,996	0.16[a]	0.05[a]
Dog	Whole-genome shotgun	215	0.05[a]	0.04[a]
Horse	Whole-genome shotgun	2,240	0.00[a]	0.00[a]
Cow	Whole-genome shotgun	442	1.40[a]	0.93[a]

Human data reflect our revised, SHIMS-based assembly.
[a]Numbers are based on whole-genome shotgun assemblies, which likely underestimate X-ampliconic content.

11.5 Mb of sequence generated, 3.15 Mb comprised X-chromosomal amplicons. We estimated the total size of the human X chromosome to be about 155.3 Mb, of which ~2% was ampliconic (**Table 1**).

Our SHIMS assembly substantially improved on the current reference sequence (**Supplementary Table 1**). It closed four amplicon-associated gaps, corrected misassemblies of three large ampliconic regions (**Fig. 2** and **Supplementary Figs. 1** and **2**) and identified two previously unrecognized palindromic amplicons (**Supplementary Fig. 3**). As an example of the improved accuracy of this approach (**Supplementary Note**), our SHIMS-based assembly of one ampliconic region closed a gap, reduced the X-chromosome reference sequence by 236 kb and showed that an apparently complex collection of amplicons was a solitary palindrome (**Fig. 2**). This SHIMS-based assembly of X-chromosomal amplicons will be incorporated into the reference sequence of the human X chromosome.

With our more accurate assembly and corresponding recalibration of the human X chromosome's gene content, we tested Ohno's law by systematically comparing the gene contents of the human and mouse X chromosomes. Contrary to Ohno's law, 18% (144/800) of human and 23% (197/853) of mouse X-linked protein-coding genes did not have orthologs in the other species (**Fig. 3a** and **Supplementary Tables 2**–**4**). In sum, this 2-species comparison identified 341 genes that violated Ohno's law.

An exception to Ohno's law could arise through either gene loss or duplication of an ancestral X-linked gene or through independent acquisition of a novel gene. To identify cases of gene loss, we searched three outgroup species for orthologs of human and mouse X-linked genes that violated Ohno's law: dog (X chromosome)[25], horse (X chromosome)[26] and chicken (in which autosomes 1 and 4 are homologous to the mammalian X chromosome)[27]. We concluded that a minority of the genes (55/144 in humans and 34/197 in mice) that violated Ohno's law were the result of lineage-specific gene loss (**Fig. 3a** and **Supplementary Tables 3** and **4**). To identify cases of duplication of an ancestral X-linked gene, we also analyzed orthologous genes in dog, horse and chicken for comparison. Only a small fraction of the genes (13/144 in humans and 29/197 in mice) that violated Ohno's law were present because of duplication of an ancestral X-linked gene (**Fig. 3a** and **Supplementary Tables 3** and **4**). These findings indicate that, in both lineages, the majority of genes (76/144 in humans and 134/197 in mice) that violated Ohno's law were independently acquired—through transposition or retroposition from autosomes or through having arisen *de novo* on the X chromosome. Thus, unexpectedly large fractions of X-linked genes (10% in humans and 16% in mice) have been acquired independently since the two lineages began to diverge from a common ancestor 80 million years ago.

We then counted the numbers of independently acquired and shared genes that were ampliconic (embedded in duplicated segments of >10 kb in length and exhibiting >99% nucleotide identity), multicopy (only the gene structure was duplicated) or single copy. Among independently acquired X-linked genes, roughly two-thirds were ampliconic (48/76 in humans and 102/134 in mice), whereas the remaining one-third were multicopy or single copy (**Fig. 3b**

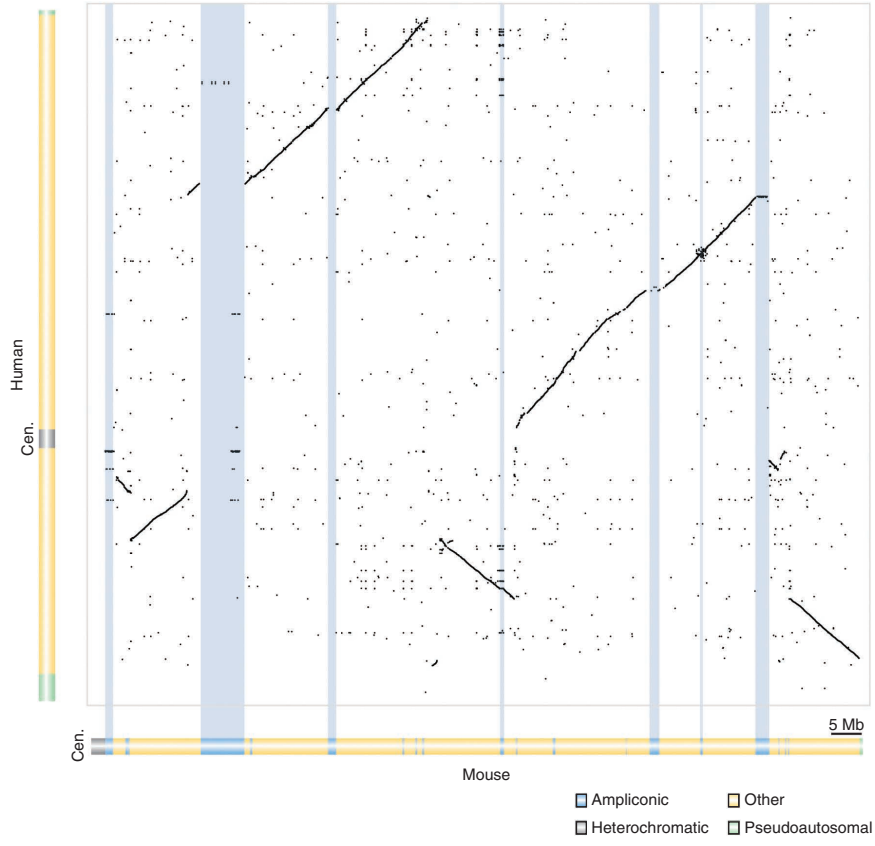


Figure 1 A dot-plot comparison of the nucleotide sequences of the human and mouse X chromosomes shows large, divergent ampliconic regions on the mouse X chromosome. The dot plot was generated from BLASTZ nucleotide alignments of the human X chromosome assembly, before our SHIMS-based refinement (*y* axis), and the single-haplotype mouse X-chromosome assembly (*x* axis); each dot represents >70% nucleotide identity within a 10-kb window centered on that position. Within the plot, diagonal lines indicate syntenic blocks between the two chromosomes; regions lacking these diagonal lines comprise species-specific sequences. Blue shading highlights divergent ampliconic regions, each >500 kb in length, on the mouse X chromosome. Cen., centromere.

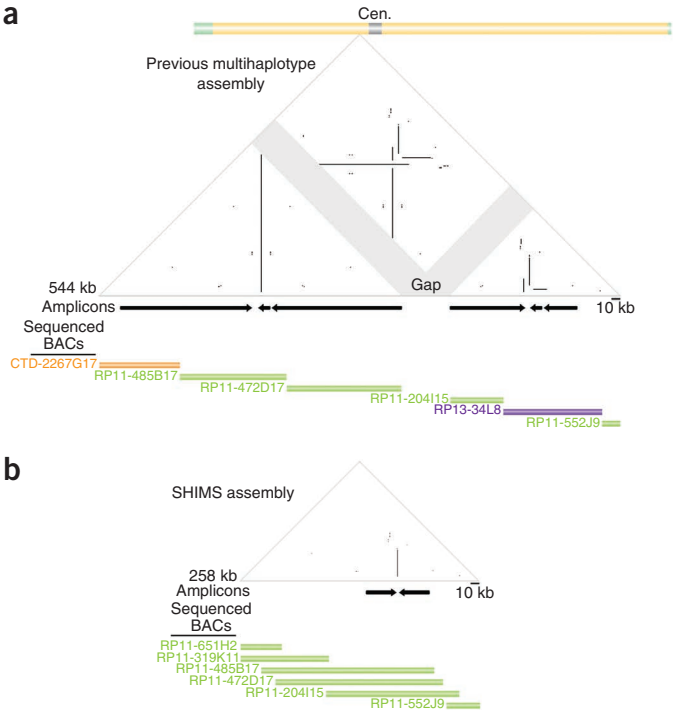
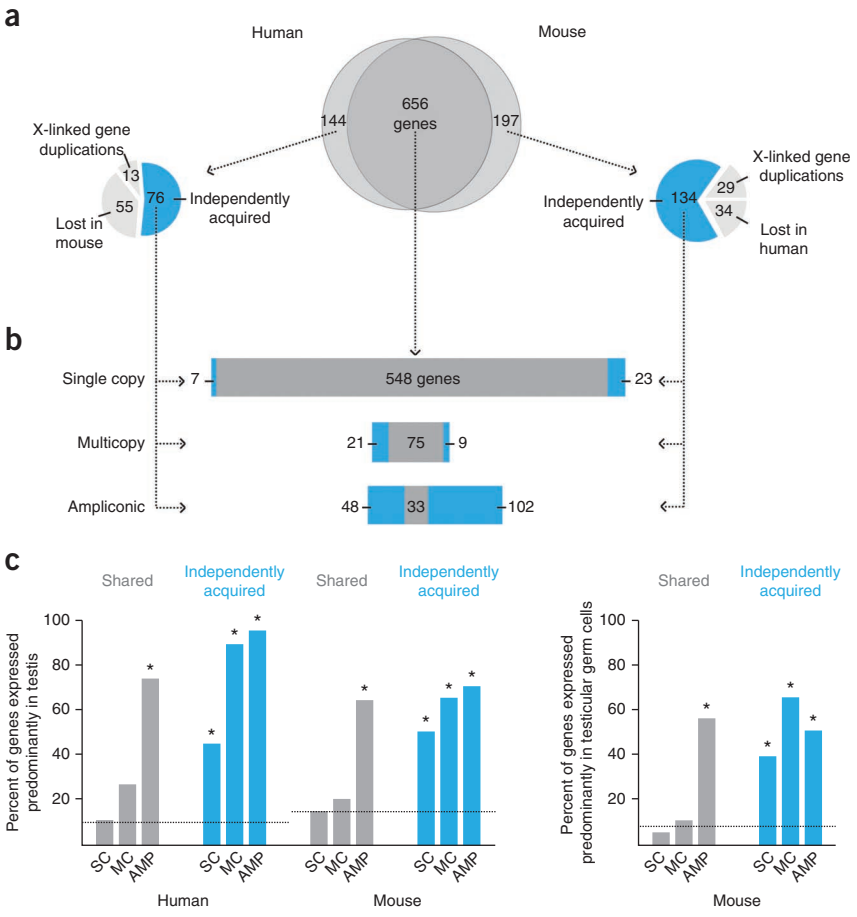


Figure 2 Comparison of mosaic and SHIMS-based sequence assemblies across one region of the human X chromosome. (**a**) The triangular dot plot highlights sequence similarities within the mosaic (multihaplotype) assembly. Each dot represents 100% identity within a window of 100 nucleotides. Direct repeats appear as horizontal lines, inverted repeats appear as vertical lines, and palindromes appear as vertical lines that nearly intersect the baseline; gaps are indicated by gray shading. Black arrows immediately below denote the positions and orientations of amplicons. Further below, sequenced BACs from CTD, RP-11 and RP-13 libraries (each from a different individual) contributing to the assembly are depicted as orange, green and purple bars, respectively; each bar reflects the extent and position within the assembly of finished sequence for that BAC. (As per the standard for the human genome assembly, finished-sequence overlaps between adjoining BACs are limited to 2 kb.) GenBank accessions are given in **Supplementary Table 1**. (**b**) SHIMS-based assembly of the same region. All BACs are derived from the RP-11 library (one male) and are fully sequenced; the finished sequence of each BAC extensively overlaps those of adjoining BACs.

and **Supplementary Tables 3** and **4**). Indeed, only 31% of human X-ampliconic genes (33/107) and 22% of mouse X-ampliconic genes (33/149) had orthologs in the other species (**Supplementary Table 5**). In contrast, 82% of shared X-linked genes were single copy (548/656; **Fig. 3b**), and an impressive 95% of human (548/575) and 94% of mouse (548/585) single-copy X-linked genes were shared (**Supplementary Table 5**). We conclude that, when comparing the X-linked genes of humans with those of mice, most exceptions to Ohno's law are ampliconic genes that were independently acquired in either the human or mouse lineage subsequent to their divergence from a common ancestor 80 million years ago (**Fig. 3b**). These exceptions provide a notable contrast to the shared, single-copy genes that follow Ohno's law.



We then compared the expression patterns of independently acquired and shared X-linked genes in eight human tissues and three mouse tissues, using published[28–30] and newly generated RNA deep sequencing (mRNA-seq) data. As a control, we analyzed all autosomal genes. We observed that most independently acquired human and mouse X-linked genes exhibited high expression in the testis and little or no expression in all other tissues examined (**Fig. 3c**, **Supplementary Fig. 4** and **Supplementary Tables 6–8**). Because many of the independently acquired genes are members of multicopy or ampliconic gene families whose gene expression levels

Figure 3 Comparison of X-linked gene classes between humans and mice. (**a**) Center, the Venn diagram depicts all human and mouse X-linked genes that are shared or not shared. Left and right, pie charts depict species-specific genes independently acquired in that lineage (blue), duplicated from an ancestral X-linked gene in that lineage (light gray) or lost in the opposite lineage (light gray). The Venn diagram and pie charts are drawn to scale (by gene number). (**b**) Horizontal bar stacks of single-copy, multicopy and ampliconic genes shared (dark gray) and independently acquired (blue) on the human and mouse X chromosomes. Bar stacks are to scale (by gene number). (**c**) Percentages of genes expressed predominantly in the testis and in testicular germ cells. Horizontal dotted lines represent the percentages of autosomal genes exhibiting testis- or testicular germ cell–predominant expression. SC, single copy; MC, multicopy; AMP, ampliconic. Each asterisk indicates χ^2 test with Yates' correction, $P < 0.0001$ (1 degree of freedom), compared to either autosomal genes or X-linked single-copy genes.

were averaged, it was important to rule out the possibility that only one family member was actively transcribed in the testis—which we did by scrutinizing the testis mRNA-seq data for sequence variants that differentiated members of a gene family (**Supplementary Table 9**). The testis-predominant expression pattern of independently acquired genes was significantly different (χ^2, $P < 0.0001$) from that of the shared, single-copy X-linked genes (**Fig. 3c** and **Supplementary Table 6**). Notably, the proportion of shared, single-copy X-linked genes that were expressed predominantly in the testis was much lower and was approximately the same as for autosomal genes (**Fig. 3c** and **Supplementary Tables 6**, **10** and **11**). In summary, we find that a common and distinguishing characteristic of most independently acquired X-linked genes is testis-predominant expression.

We next sought to determine whether independently acquired X-linked genes in mice are expressed in germ cells or somatic cells of the testis. To do this, we performed mRNA-seq analysis on adult testes from wild-type and Kit^{W}/Kit^{Wv} mice, the latter of which lack germ cells[31]. We found that most independently acquired genes were expressed specifically in testicular germ cells, regardless of whether they were single copy, multicopy or ampliconic (**Fig. 3c** and **Supplementary Tables 6–8**). The proportion of independently acquired genes with high expression in wild-type testis and little or no expression in Kit^{W}/Kit^{Wv} testis was significantly higher (χ^2, $P < 0.0001$) than that of either shared single-copy X-linked genes or autosomal genes (**Fig. 3c** and **Supplementary Tables 6**, **8** and **11**). Additionally, in accordance with our previous studies[32], we found that most ampliconic genes, both shared and independently acquired, were also predominantly expressed in testicular germ cells (**Fig. 3c** and **Supplementary Tables 6** and **8**). Our findings underscore the importance of the male germ line, relative to the soma, in promoting gene acquisition on a chromosome whose gene content is otherwise highly conserved.

On the basis of our present findings in humans and mice, we wonder whether the X chromosomes of other placental mammals (**Supplementary Fig. 5**) also harbor independently acquired ampliconic genes that are expressed predominantly in testicular germ cells. To answer this question in other species will require using a SHIMS approach to assemble their X amplicons and, thus, their reference sequences completely and accurately (**Table 1**). If independently acquired, testis-expressed genes prove to be a general feature of mammalian X chromosomes, then the acquisition of these genes may have contributed greatly to mammalian diversification and radiation, which began in the Paleocene epoch. This speculation is supported by a wealth of evidence that the rapid evolution of hybrid male-sterility factors on animal X chromosomes has been a key driver of speciation[33]. In mice, three strictly X-linked hybrid male-sterility loci[34–36] have been identified; all three map at or near X-ampliconic regions (**Supplementary Fig. 6**) harboring independently acquired genes expressed predominantly in spermatogenic cells.

The medical and biological significance of the independently acquired genes on the human X chromosome is essentially unexplored. So far, not a single X-linked phenotype (as cataloged in Online Mendelian Inheritance in Man, OMIM) has been attributed at the molecular level to an independently acquired gene on the human X chromosome (**Supplementary Table 12**). In contrast, 238 X-linked traits have been traced at the molecular level to genes shared by humans and mice. Given that the independently acquired genes are expressed predominantly in spermatogenic cells, one might anticipate that loss-of-function mutations affecting these genes or gene families would perturb male gametogenesis—a possibility that can now be explored using the SHIMS-based reference sequence of the human X-ampliconic regions.

Our findings also provide a plausible explanation for how so many X-linked genes are able to defy Ohno's law. Ohno's law assumed that any given X-linked gene would be expressed in both sexes and equally so. Consistent with this idea, we found that, in both humans and mice, >96% of genes that followed Ohno's law were expressed in both sexes (**Supplementary Tables 13** and **14**). However, not all genes function equivalently in males and females, and, indeed, some genes were expressed in one sex but not in the other. As we have shown, the genes that violated Ohno's law were expressed in males but not in females. The fact that many genes are expressed in a sex-specific manner would not have been appreciated at the time of Ohno's writing in the 1960s.

In summary, our study places Ohno's law within a larger context. Based on the construction and analysis of a more complete and accurate human X-chromosome reference sequence, our comparison of human and mouse X chromosomes enables us to characterize key exceptions to the law: in both species, large numbers of genes are expressed in spermatogenic cells, most of which are ampliconic or multicopy. We conclude that the gene repertoires of the human and mouse X chromosomes are products of two complementary evolutionary processes: conservation of single-copy genes that serve in functions shared by the sexes and ongoing gene acquisition, usually involving the formation of amplicons, which leads to the differentiation and specialization of X chromosomes for functions in male gametogenesis.

URLs. LASTZ, http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/README.lastz-1.02.00a.html; R software for dot plots, http://www.r-project.org/; custom Perl script for triangular dot plots, http://pagelab.wi.mit.edu/material-request.html.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Kit^{W}/Kit^{Wv} and Kit^{+}/Kit^{Wv} testis mRNA-seq reads have been deposited in GenBank under accession SRA060831. SHIMS-based assemblies have been deposited in GenBank under accessions JH720451–JH720454, JH806587–JH806603, JH159150, KB021648. Specific information about the BACs and fosmids used to generate the SHIMS-based assemblies is provided in **Supplementary Table 1**.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D. Albracht, J. Collins, M. Gill, N. Koutseva, C. Kremitzki, A. van der Veen and J. Wood for technical assistance and D. Bellott, R. Desgraz, G. Dokshin, T. Endo, A. Godfrey, Y. Hu, J. Hughes, M. Kojima, B. Lesch, L. Okumura, K. Romer and Y. Soh for comments on the manuscript. This work was supported by the US National Institutes of Health and the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

J.L.M., H.S., W.C.W., R.K.W. and D.C.P. planned the project. J.L.M. and L.G.B. performed BAC mapping. J.L.M. performed RNA deep sequencing. T.G., S.R., K.A. and S.Z. were responsible for finished BAC sequencing. J.L.M. and H.S. performed sequence analyses. J.L.M. and D.C.P. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Ohno, S. *Sex Chromosomes and Sex-Linked Genes* (Springer, Berlin, 1967).
2. Kuroiwa, A. *et al.* Conservation of the rat X chromosome gene order in rodent species. *Chromosome Res.* **9**, 61–67 (2001).

3. Delgado, C.L., Waters, P.D., Gilbert, C., Robinson, T.J. & Graves, J.A. Physical mapping of the elephant X chromosome: conservation of gene order over 105 million years. *Chromosome Res.* **17**, 917–926 (2009).
4. Prakash, B., Kuosku, V., Olsaker, I., Gustavsson, I. & Chowdhary, B.P. Comparative FISH mapping of bovine cosmids to reindeer chromosomes demonstrates conservation of the X-chromosome. *Chromosome Res.* **4**, 214–217 (1996).
5. Ross, M.T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434**, 325–337 (2005).
6. Veyrunes, F. *et al.* Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Res.* **18**, 965–973 (2008).
7. Watanabe, T.K. *et al.* A radiation hybrid map of the rat genome containing 5,255 markers. *Nat. Genet.* **22**, 27–36 (1999).
8. Raudsepp, T. *et al.* Exceptional conservation of horse-human gene order on X chromosome revealed by high-resolution radiation hybrid mapping. *Proc. Natl. Acad. Sci. USA* **101**, 2386–2391 (2004).
9. Band, M.R. *et al.* An ordered comparative map of the cattle and human genomes. *Genome Res.* **10**, 1359–1368 (2000).
10. Murphy, W.J., Sun, S., Chen, Z.Q., Pecon-Slattery, J. & O'Brien, S.J. Extensive conservation of sex chromosome organization between cat and human revealed by parallel radiation hybrid mapping. *Genome Res.* **9**, 1223–1230 (1999).
11. Spriggs, H.F. *et al.* Construction and integration of radiation-hybrid and cytogenetic maps of dog chromosome X. *Mamm. Genome* **14**, 214–221 (2003).
12. Palmer, S., Perry, J. & Ashworth, A. A contravention of Ohno's law in mice. *Nat. Genet.* **10**, 472–476 (1995).
13. Rugarli, E.I. *et al.* Different chromosomal localization of the *Clcn4* gene in *Mus spretus* and C57BL/6J mice. *Nat. Genet.* **10**, 466–471 (1995).
14. She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930 (2004).
15. Olivier, M. *et al.* A high-resolution radiation hybrid map of the human genome draft sequence. *Science* **291**, 1298–1302 (2001).
16. Dietrich, W.F. *et al.* A comprehensive genetic map of the mouse genome. *Nature* **380**, 149–152 (1996).
17. Church, D.M. *et al.* Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.* **7**, e1000112 (2009).
18. Tishkoff, S.A. & Kidd, K.K. Implications of biogeography of human populations for 'race' and medicine. *Nat. Genet.* **36**, S21–S27 (2004).
19. Bovee, D. *et al.* Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nat. Genet.* **40**, 96–101 (2008).
20. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
21. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
22. Hughes, J.F. *et al.* Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**, 536–539 (2010).
23. Kuroda-Kawaguchi, T. *et al.* The *AZFc* region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nat. Genet.* **29**, 279–286 (2001).
24. Bellott, D.W. *et al.* Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* **466**, 612–616 (2010).
25. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
26. Wade, C.M. *et al.* Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326**, 865–867 (2009).
27. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
28. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
29. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
30. Bradley, R.K., Merkin, J., Lambert, N.J. & Burge, C.B. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. *PLoS Biol.* **10**, e1001229 (2012).
31. Handel, M.A. & Eppig, J.J. Sertoli cell differentiation in the testes of mice genetically deficient in germ cells. *Biol. Reprod.* **20**, 1031–1038 (1979).
32. Mueller, J.L. *et al.* The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. *Nat. Genet.* **40**, 794–799 (2008).
33. Coyne, J.A. & Orr, H.A. *Speciation* (Sinauer Associates, Sunderland, MA, 2004).
34. Elliott, R.W. *et al.* Genetic analysis of testis weight and fertility in an interspecies hybrid congenic strain for chromosome X. *Mamm. Genome* **12**, 45–51 (2001).
35. Elliott, R.W., Poslinski, D., Tabaczynski, D., Hohman, C. & Pazik, J. Loci affecting male fertility in hybrids between *Mus macedonicus* and C57BL/6. *Mamm. Genome* **15**, 704–710 (2004).
36. Storchová, R. *et al.* Genetic analysis of X-linked hybrid sterility in the house mouse. *Mamm. Genome* **15**, 515–524 (2004).

ONLINE METHODS

Nucleotide sequence comparisons. Entire X-chromosome files for human (hg18), chimpanzee (panTro4), rhesus (rheMac3), mouse (mm9), dog (canFam3), horse (equCab2) and cow (bosTau7) were downloaded from the UCSC Genome Browser database[37]. Alignments of repeat-masked X-chromosome sequences were generated with BLASTZ[38], using non-gapped alignment settings and a step length of 20 nucleotides. Coordinates were obtained for stretches of alignable sequence that scored >3,000 (the equivalent of a 30-bp perfect match), using default gap and mismatch penalty and reward parameters. Square dot plots were generated with the R software dot-plot package. Triangular dot plots were generated using a custom Perl script[23].

Selection of human X-chromosome regions for single-haplotype assembly. To identify human X-ampliconic regions, we considered regions of the human X-chromosome reference sequence falling into one of three categories.

(1) Regions containing amplicons in the current reference sequence. A collection of previously detected segmental duplications[39] was filtered for duplications meeting the following criteria: repeat unit of >10 kb with >99.0% identity between copies and <500 kb of separation between copies.

(2) Regions containing gaps. Ampliconic sequences are known to be associated with gaps in genome assemblies[14]. In the human reference sequence, gaps are marked by long stretches of Ns, denoting missing sequence of unknown length. We scanned the non-pseudoautosomal reference sequence (hg18) for such large stretches of Ns and identified the sequence coordinates of nine gaps.

(3) Regions containing misoriented, physically mapped clones. We used position and orientation information for fosmid paired-end sequences (from eight different libraries[20]) previously aligned to the human X-chromosome reference sequence. X-chromosome regions where fosmid paired-end sequences from at least three libraries did not map in similar orientation to the reference sequence were considered putative ampliconic regions.

A total of 33 regions (**Supplementary Table 1**) were identified using these 3 approaches. For each of the 33 regions, we identified the library of origin for each reference sequence clone to determine which portions of the reference assembly, if any, were composed of single-haplotype sequence.

Clone selection and sequencing. For the 29 regions not comprising single-haplotype sequence, we employed the single-haplotype iterative mapping and sequencing (SHIMS) approach, as previously performed for Y- and Z-chromosome assemblies[21–24]. We used publicly available BAC fingerprint maps, fosmid and BAC end sequences and current X-chromosome reference sequence[5] as sources for generating markers.

We selected and sequenced BACs and fosmids that collectively spanned each region. For each region that included a gap in the current reference sequence, we selected a tiling path of clones stretching 500 kb on either side of the gap. Analysis of this ~1-Mb segment of sequence allowed us to determine whether sequences flanking the gap were ampliconic. For each region of the X chromosome that appeared to be ampliconic in the current reference sequence or that contained misoriented fosmid ends, we selected a tiling path of clones stretching 100 kb on either side of the amplicon or misoriented fosmid end sequence.

We primarily selected human X-chromosome BACs from the RP-11 male library[40]. In those instances where RP-11 BACs did not provide sufficient coverage of a region, we selected clones from the haploid CH-17 library. In some instances, amplicon repeat units were too short to be assembled accurately within a single BAC (average BAC insert size of 160 kb). For such cases, we selected clones from the ABC8 male fosmid library[20]; these clones have smaller inserts (~40 kb), which enabled us to order and orient amplicons with shorter repeat units. We used only one library (either RP-11, CH-17 or ABC-8) to span each of the 29 ampliconic regions sequenced. In a few cases, we used alternative ABC fosmids (from libraries ABC-7, ABC-9, ABC-12, ABC-13 and ABC-14)[20] to extend into gaps that were not ampliconic.

BAC and fosmid sequences will be incorporated into the next update of the reference assembly (GRCh38). GenBank accessions for all BACs and fosmids as well as for SHIMS-based assemblies of the 29 regions sequenced are provided in **Supplementary Table 1**.

Comparisons of human and mouse gene orthologs. Reference sequences for the human (hg18) and mouse (mm9) protein-coding gene sets were downloaded from the UCSC Genome Browser[37]. We selected the isoform yielding the longest peptide sequence for each gene, resulting in 821 and 865 genes for the human and mouse X chromosomes, respectively. These lists of genes were curated to provide an unbiased and comprehensive comparison of human and mouse X-linked gene content, as follows.

(1) All pseudoautosomal genes were removed because our analysis was limited to strictly X-linked genes. The case of the *STS* gene (encoding steroid sulfatase) merits special mention. The human *STS* gene is X linked. In mice, *Sts* is absent from the reference genome assembly, but multiple ESTs have been reported. Previous studies[41] and our data (S. Soh and H.S., unpublished data; see **Supplementary Table 2** for more information) are consistent with the mouse *Sts* gene mapping to the X chromosome, in or near the pseudoautosomal region. We included *Sts* in the mouse gene set.

(2) For 11 genes (**Supplementary Table 2**), we determined that the gene was multicopy in humans but ampliconic in mice or vice-versa. We excluded these genes from all tallies and analyses because we could not infer whether the genes were multicopy or ampliconic in the common ancestor of humans and mice.

(3) We updated and corrected the human gene set to reflect our SHIMS sequence assembly across ampliconic regions. We searched novel genomic sequence generated in this study for genes using Genomescan[42] and BLAST[43] analyses of human EST databases. In the case of ampliconic regions that were either expanded or contracted in our revised assembly, we recounted the numbers of genes for each gene family within the regions.

Through this curation, we arrived at a total of 800 human and 853 mouse X-linked genes. These revised gene sets served as the basis for all subsequent comparative and expression analyses.

All X-linked genes determined to be shared by humans and mice were identified by having either a best reciprocal BLAST[43] alignment in the two species or a TBLASTN alignment to a syntenic, unannotated region of the compared X chromosome (**Supplementary Table 2**). Such regions were classified as unannotated genes when the predicted protein-coding gene sequence was free of nonsense mutations and there was evidence of transcription from either EST or mRNA-seq data[28,29].

Genes present on either the human or mouse X chromosome (but not both) could either have been lost in one lineage, duplicated in one lineage or independently acquired in one lineage. To distinguish between these three possibilities, we determined, via TBLASTN, whether X-linked genes present in either humans or mice (but not both) had orthologs on the dog X chromosome (canFam3), the horse X chromosome (equCab2) or syntenic regions of chicken chromosomes 1 and 4 (galGal4). Comparisons with these three outgroups helped us to infer whether a given gene was present on the X chromosome in the common ancestor of humans and mice. Each gene was classified as follows.

(1) Lineage-specific gene loss. A gene with an ortholog in a syntenic chromosomal region in one or more of the three outgroups or with a pseudogene ortholog in the syntenic region of the human or mouse X chromosome was judged to have been lost.

(2) Lineage-specific gene duplication. A gene duplicate (paralog) of a pre-existing X-linked gene that did not have an orthologous duplicate gene in the other species (human or mouse) or in a syntenic chromosomal region in one or more of the three outgroups was judged to be a lineage-specific duplicate of a pre-existing X-linked gene.

(3) Independently acquired. A gene not falling into either of these two categories was judged to have been independently acquired.

Our inferences regarding human and mouse X-linked gene losses and gains were based on comparisons with the current dog, horse and chicken genome assemblies. As the assemblies of the dog, horse and chicken genomes are not as complete as those of the human and mouse X chromosomes, our inferences should be re-examined in the future when more complete and accurate assemblies of the dog X chromosome, horse X chromosome and chicken chromosomes 1 and 4 are available. In the **Supplementary Note**, we elaborate on these limitations and the associated uncertainties.

Shared and species-specific genes were grouped as single copy, multicopy or ampliconic. We defined multicopy genes as members of gene pairs or families exhibiting >50% amino acid identity across 80% of the protein and

doi:10.1038/ng.2705

with an *e* value of $<1 \times 10^{-20}$ when protein sequences were aligned[44]. We defined ampliconic genes as genes located within a stretch of ampliconic sequence (repeat unit of >10 kb in length with >99% nucleotide identity and <500 kb of separation between repeat units).

mRNA-seq of testis cDNA. We crossed C57BL/6J Kit^{Wv} (The Jackson Laboratory) males to WB/ReJ Kit^{W} (The Jackson Laboratory) females to generate Kit^{W}/Kit^{Wv} compound-heterozygous males, which are germ cell deficient, and control Kit^{+}/Kit^{Wv} males. Two biological replicate testes from Kit^{W}/Kit^{Wv} and Kit^{+}/Kit^{Wv} males were collected at ~3 months of age. Total RNA (1–2 µg) was extracted using TRIzol (Invitrogen) according to the manufacturer's instructions. Hemoglobin transcripts were selectively removed from total RNA by following the recommendations in the GLOBINclear (Ambion) protocol. Per the Illumina mRNA-Seq sample prep kit protocol, polyA-selected mRNA was used to generate mRNA-seq cDNA libraries using random-hexamer primers. cDNA fragments of ~200 nucleotides were isolated and modified for sequencing following the mRNA-seq protocol (Illumina). The Illumina Genome Analyzer II platform was used to sequence 36-mers from the mRNA-seq libraries following the manufacturer's recommendations. Kit^{W}/Kit^{Wv} and Kit^{+}/Kit^{Wv} testis mRNA-seq reads have been deposited in GenBank under accession SRA060831. The Massachusetts Institute of Technology's committee on animal care has approved all experiments involving mice.

RNA-seq analyses. Previously published mRNA-seq data sets from human[28,30] (adipose, colon, heart, liver, lymph node, skeletal muscle, ovary and testis) and mouse[29] (liver and skeletal muscle) tissues were combined with our newly generated Kit^{+}/Kit^{Wv} testis and Kit^{W}/Kit^{Wv} testis data sets to determine the tissue expression pattern for each X-linked gene. For each tissue, mRNA-seq reads were aligned to the reference genome assembly using Tophat[45] with default settings. FPKM (fragments per kilobase of exon model per million mapped fragments) values were estimated using Cufflinks[46] with the reference sequence gene set used as an annotation file. Unannotated genes with orthologs in the reciprocal species (29 such cases) were excluded owing to concerns regarding accurate estimates of FPKM.

Cufflinks has difficulty accurately calculating FPKM values for multicopy and ampliconic genes, so we estimated FPKM values for these two gene classes using a customized method. FPKM values for multicopy and ampliconic genes were determined by aligning all reads to a representative gene family member. This total read count per gene family was then divided by the length of the gene, the number of gene copies and the number of reads mapped to the genome, resulting in an FPKM value for each ampliconic or multicopy gene family. To determine whether multiple members of a multicopy or ampliconic gene family were expressed, we identified nucleotide variants that uniquely identified individual copies. We then counted the number of mRNA-seq reads in human[30] and mouse testis samples[47] that aligned to each variant.

Genes with an FPKM value of >1 in testis and <1 in ovaries and all somatic tissues examined were considered to be expressed predominantly in the testis. Similarly, genes with an FPKM value of >1 in Kit^{+}/Kit^{Wv} testes and <1 in Kit^{W}/Kit^{Wv} testis and in all other somatic tissues examined were considered to be expressed predominantly in testicular germ cells. (Previous studies have used FPKM of >1 as a cutoff for considering a gene to be expressed in a tissue[48].)

To determine whether X-linked genes that followed Ohno's law were expressed in both sexes, we analyzed previously published mRNA-seq data sets from male and female human and mouse tissues[47]. We performed alignments to calculate FPKM values as described. We considered a gene to be expressed in one sex but not in the other if it met both of the following criteria: (i) FPKM of >1 in one sex and FPKM of <1 in the other sex and (ii) at least threefold higher expression in one sex compared with the other sex.

37. Fujita, P.A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* **39**, D876–D882 (2011).
38. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
39. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
40. Osoegawa, K. *et al.* A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.* **11**, 483–496 (2001).
41. Salido, E.C. *et al.* Cloning and expression of the mouse pseudoautosomal steroid sulphatase gene (*Sts*). *Nat. Genet.* **13**, 83–86 (1996).
42. Yeh, R.F., Lim, L.P. & Burge, C.B. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11**, 803–816 (2001).
43. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
44. Thornton, K. & Long, M. Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. *Mol. Biol. Evol.* **19**, 918–925 (2002).
45. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
46. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
47. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).
48. Deng, X. *et al.* Evidence for compensatory upregulation of expressed X-linked genes in mammals, *Caenorhabditis elegans* and *Drosophila melanogaster*. *Nat. Genet.* **43**, 1179–1185 (2011).